FateZero: Fusing Attentions for Zero-shot Text-based Video Editing

Year: 2023 | Citations: 452 | Authors: Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang

---

Abstract

The diffusion-based generative models have achieved remarkable success in text-based image generation. However, since it contains enormous randomness in generation progress, it is still challenging to apply such models for real-world visual content editing, especially in videos. In this paper, we propose FateZero, a zero-shot text-based editing method on real-world videos without per-prompt training or use-specific mask. To edit videos consistently, we propose several techniques based on the pre-trained models. Firstly, in contrast to the straightforward DDIM inversion technique, our approach captures intermediate attention maps during inversion, which effectively retain both structural and motion information. These maps are directly fused in the editing process rather than generated during denoising. To further minimize semantic leakage of the source video, we then fuse self-attentions with a blending mask obtained by cross-attention features from the source prompt. Furthermore, we have implemented a reform of the self-attention mechanism in denoising UNet by introducing spatial-temporal attention to ensure frame consistency. Yet succinct, our method is the first one to show the ability of zero-shot text-driven video style and local attribute editing from the trained text-to-image model. We also have a better zero-shot shape-aware editing ability based on the text-to-video model [52]. Extensive experiments demonstrate our superior temporal consistency and editing capability than previous works.